Ensemble Data Assimilation: Theory and Applications

Humberto C. Godinez

J. David Moulton, Jon Reisner, Alex O. Fierro, Stephen Guimond, Jim Kao

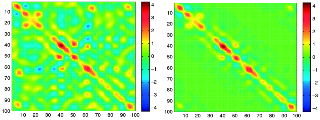
Los Alamos National Laboratory

National Severe Storms Laboratory Seminar
July 6, 2011

Outline

1. Introduction
2. Ensemble Data Assimilation
   - Matrix-Free EnKF
   - Sherman-Morrison-Woodbury solver
   - Comparison Experiments
3. Hurricane Data Assimilation
   - Hurricane Model
   - Dual-doppler Radar Data
4. Assimilation Results
   - Model and Ensemble Setup
   - Ensemble Structure
   - Parameter Estimation
   - Simulation Experiments
5. Conclusion and Future Work
   - extra

Introduction

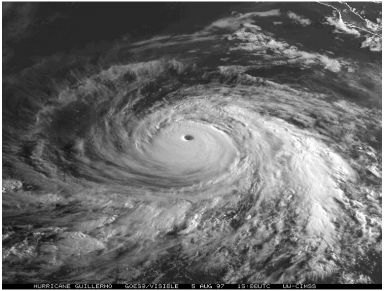


★

- Ensemble data assimilation widely in weather and climate community.
- Ensemble assimilation has various challenges.
- Theoretical: ensemble initialization, covariance localization, inflation, model error, covariance sensitivity, etc..
- Computational: ensemble simulation, assimilation of large data sets, high dimensional model simulation, etc..

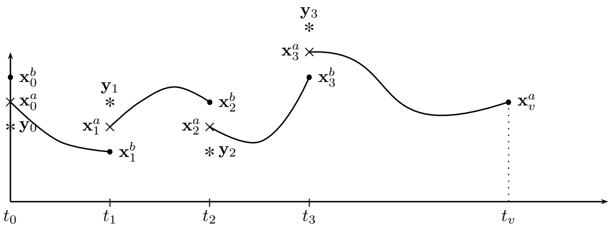
★Petrie, Ruth E. and Dance, Sarah L., *Ensemble-based data assimilation and the localisation problem*, Weather, **65**, pp 65–69, 2010.

Introduction cont.



- Hurricanes are extreme weather events with a high impact on society
- Accurate hurricane simulation is challenging: many physical processes, time and space scales, complexity.
- Assimilation for hurricane simulation and prediction is an active area of research, lots of work to be done!

Ensemble Data Assimilation



Data Assimilation: Methods to produce an accurate estimate of the state of a model for a given data set (observations).
Ensemble Kalman Filter (EnKF): Sequential data assimilation method that uses an ensemble of model states to calculate the state mean and error covariance matrix needed to compute an improved model state.

Matrix-Free EnKF

EnKF equations:

$$\begin{aligned} \mathbf{x}_i^a &= \mathbf{x}_i^f + \mathbf{K}\left(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^f\right), \\ \mathbf{K} &= \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1} \end{aligned}$$

rewrite as

$$\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)\mathbf{z}_i = \left(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^f\right)$$

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{P}^f\mathbf{H}^T\mathbf{z}_i$$

- Typical techniques use SVD or Cholesky decomposition.
- For large “assimilation systems”, matrix operations become expensive and matrices may be too large to hold in memory. (dense set of observations)
- EnKF can be done matrix-free.
- Iterative methods deemed to expensive to solve the linear system.

Sherman-Morrison-Woodbury solver

$$\left(\frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{H}\mathbf{x}^f - \mathbf{H}\overline{\mathbf{x}}^f\right) \left(\mathbf{H}\mathbf{x}^f - \mathbf{H}\overline{\mathbf{x}}^f\right)^T + \mathbf{R} \right) \mathbf{z}_i = \left(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^f\right)$$

Sherman-Morrison-Woodbury identity:

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}.$$

Egidi and Maponi (2006) developed a direct solver by repeatedly applying the Sherman-Morrison-Woodbury identity to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for

$$\mathbf{A} = \mathbf{A}_0 + \sum_{i=1}^{N} \mathbf{u}_i \mathbf{v}_i^T$$

- **computational cost**: linear in the number of observations and state dimension.
- **compatibility**: Ideally suited for our linear system.
- **multiple rhs**: Proportional to the cost of a vector dot product.

2-D Shallow Water Model

A global 2D shallow water (SW) model on a sphere is used for the numerical experiments.

- Model describes hydrodynamic flow on a sphere assuming vertical motion is much smaller than horizontal motion.
- Assume fluid depth is small compared with radius of the sphere (radius of Earth).
- Computations done on a $2.5^\circ \times 2.5^\circ$ grid with a time step $\Delta t = 450\mathrm{s}$.
- $\mathbf{x}_0^t$: trajectory produced by SW integration with an initial fluid depth defined by

$$h_0(\lambda, \theta) = \frac{1}{g}\left(\bar{\Phi} + 2\Omega a \vartheta \sin^3(\theta)\cos(\theta)\sin(\lambda)\right),$$

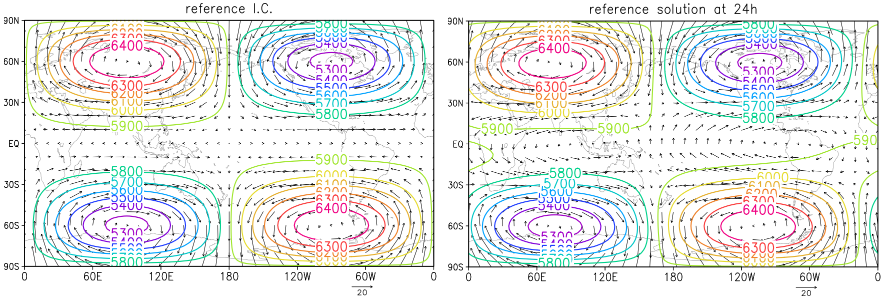
  and initial velocities u_0, v_0 derived from the geostrophic relations.
- initial condition for the control run $\mathbf{x}_0^b$ is taken from shifting $\mathbf{x}_0^t$ one grid point to the left.

Timing Experiments

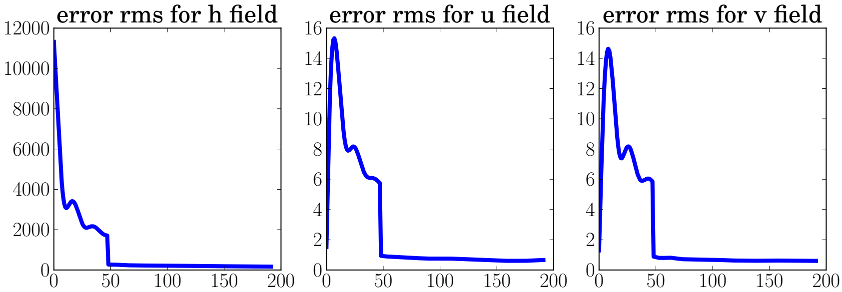
Comparison of timing between our Matrix-Free EnKF and an SVD based implementation from Evensen 2003.

- Ensemble of size $N = 100$ model simulations was used.
- Ensemble I.C. generated by perturbing $\mathbf{x}_0^b$ with a random vector sampled from a normal distribution with mean zero.
- A single assimilation was performed to compare timing between both methods.
- Ensemble is integrated up to 24h or 192 time steps.
- A single assimilation is done at 6h or time step 48.
- Varied the number of observations to assimilate from 200 to 3×10^7.
- Overhead of both methods are compared, as well as subsequent operations.

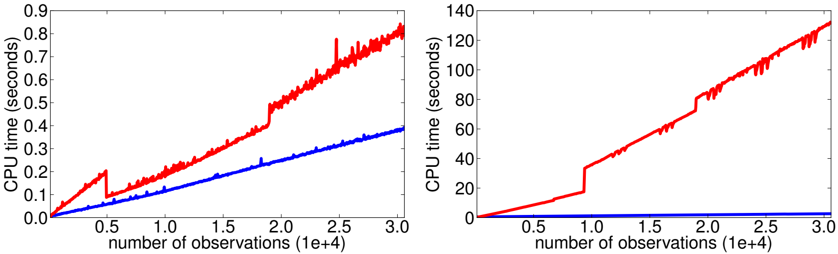
SW Reference Solution



RMS Error


error rms for h field
12000
10000
8000
6000
4000
2000
0
0
50
100
150
200
error rms for u field
16
14
12
10
8
6
4
2
0
0
50
100
150
200
error rms for v field
16
14
12
10
8
6
4
2
0
0
50
100
150
200

CPU Timing Results



Left figure: Overhead cost of SM (blue) and SVD (red). Right figure: subsequent cost of additional rhs for SM solver (blue) and matrix operations for the SVD implementation (red).

Hurricane Data Assimilation

- Hurricane track, structure, and intensification are important characteristics in simulation and prediction
- Hurricane data assimilation is an active area of research
- Several studies have used ensemble-based methods (Zhang *et.al.* 2009, MWR) and variational methods (Zou *et.al.* 2010, MWR)
- Our work concentrates in determining key model parameters to improve hurricane simulations through data assimilation

HIGRAD Hurricane Model

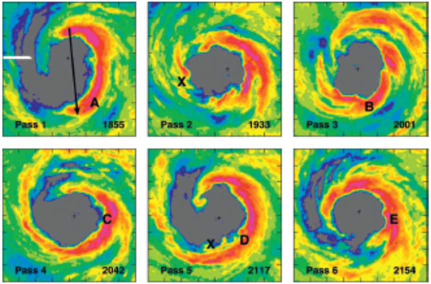
- HIgh GRADient applications model (HIGRAD) used for hurricane simulations
- HIGRAD is composed of a predictive model, based on the Navier-Stokes equation set, coupled to a bulk cloud model
- Bulk micro-physical model based on continuous approximation presented in Reisner and Jefferey (2009)
- Discretizations on A-grid using semi-implicit procedure with 4th order Runge-Kutta time discretization
- The advection scheme used was the quadratic upstream interpolation for convective kinematics advection scheme (QUICKEST, Leonard and Drummond 1995) in combination with a flux-corrected transport procedure (Zalesak 1979).

Parameters of interest

- wind shear ϕ_{shear}: tuning coefficient that determines the shear impacting the hurricane
- surface friction $\kappa_{surfacefriction}$: coefficient related to the no-slip boundary condition, magnitude impacts intensity and structure of the hurricane
- surface moisture $q_{v\,surface}$: tunning coefficient that controls the amount of surface moisture in the hurricane
- turbulent length scale ϕ_{turb}: tunning coefficient associated with turbulent transport of water vapor from the surface to the free atmosphere

Los Alamos NATIONAL LABORATORY EST. 1943

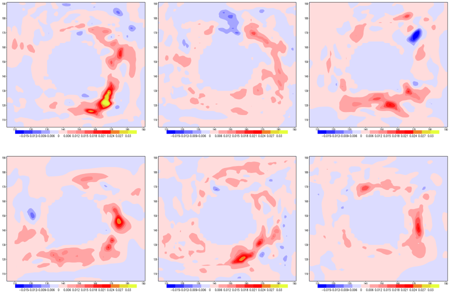
Dual-doppler data



Reasor, Paul D., Matthew D. Eastin, John F. Gamache, 2009: *Rapidly Intensifying Hurricane Guillermo (1997). Part I: Low-Wavenumber Structure and Evolution*. Mon. Wea. Rev., **137**, 603-631.

- Guillermo 1997: strong shear storm
- Two NOAA WP-3D research aircraft observed the inner core of Hurricane Guillermo from 1830 UTC 2 August to 0030 UTC 3 August
- Dual-doppler reflectivity observations were collected in 10 flight passes

Derived data fields



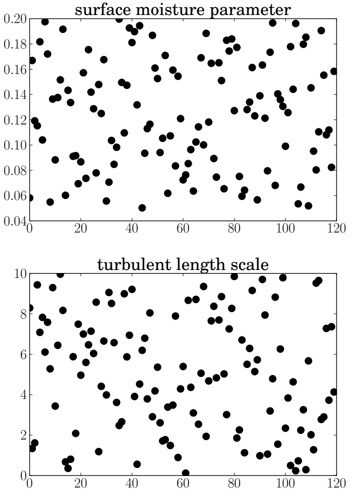
- Primary driver for a hurricane is latent heat release arising from condensation of rainwater to water vapor
- Horizontal wind fields: retrieved using a variational approach
- Latent heat fields: retrieved using derived winds and liquid water content following Guimond *et.al.* (2011, J. Atmos. Sci.)

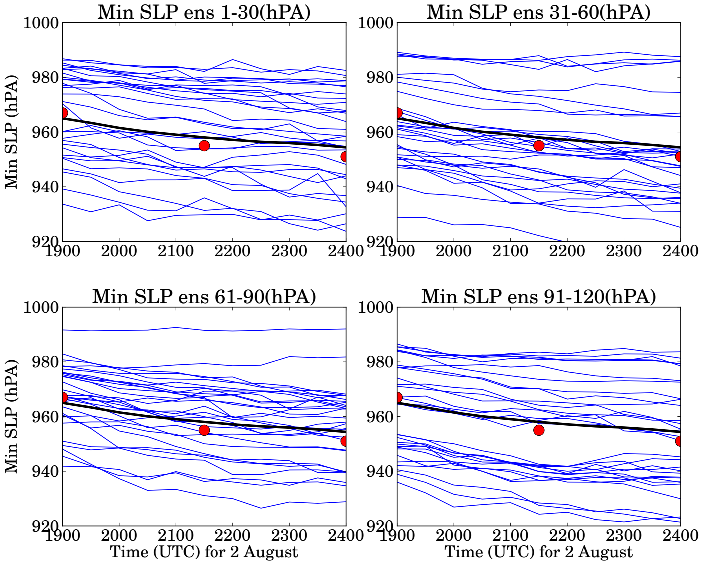
Model and Ensemble Setup

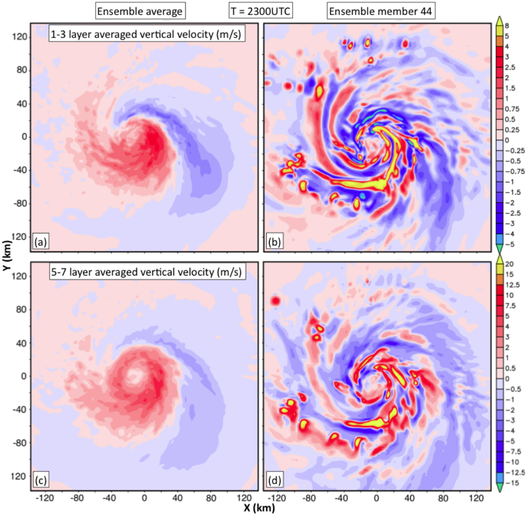
- Initialization of HIGRAD model via nudging of horizontal momentum fields and latent heat fields derived from Guillermo data
- Model spinup is run for a total of 11 hours with 6 hours for initialization and 5 hours for comparison against the dual-Doppler radar data (starting time period is 1900 UTC)
- Background average wind added to keep storm in center of discretization domain
- Stretch mesh used for horizontal and vertical spatial discretization
- Time step size was limited to 1 s to avoid any instabilities associated with exceeding the advective Courant number limit.

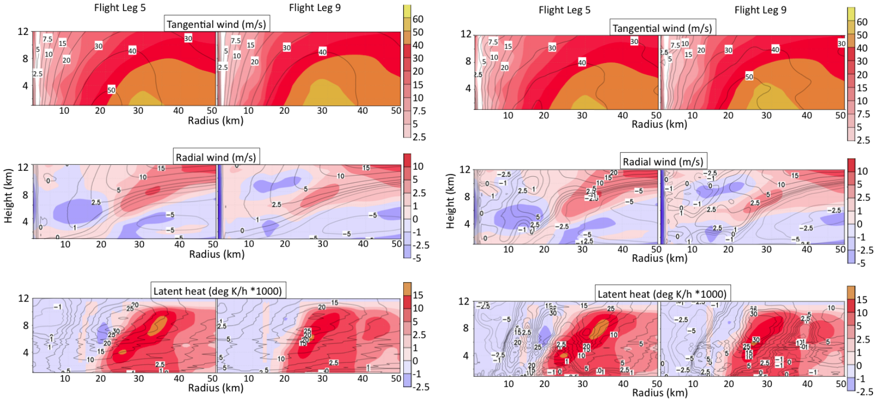
parameter	interval
surface moisture	$[0.05, 0.2]$
wind shear	$[0.1, 1.0]$
turbulent length scale	$[0.1, 10.0]$
surface friction	$[0.1, 10.0]$

- 120 ensemble members were generated by perturbing only parameters with Latin Hypercube
- Each ensemble simulation needed 225, entire ensemble typically utilized 27,000 processors on Oak Ridge National Lab's Jaguar computing platform


surface moisture parameter
wind shear parameter
turbulent length scale
surface friction parameter


Min SLP ens 1-30(hPA)
Min SLP ens 31-60(hPA)
Min SLP ens 61-90(hPA)
Min SLP ens 91-120(hPA)
Min SLP (hPA)
1000
980
960
940
920
1900
2000
2100
2200
2300
2400
Time (UTC) for 2 August


Ensemble average
T = 2300UTC
Ensemble member 44
1-3 layer averaged vertical velocity (m/s)
5-7 layer averaged vertical velocity (m/s)
(a)
(b)
(c)
(d)
Y (km)
X (km)

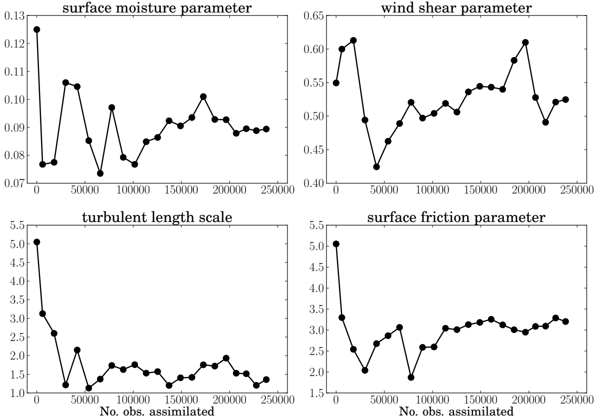


Comparisons between azimuthally-averaged profiles for the ensemble average (left) and ensemble 44 (right) and observations. Model in contours and observations are shaded. Fields are for tangential winds (top), radial winds (center), and latent heat (bottom). Time periods for comparisons are at flight leg 5 (2117 UTC) and 9 (2333 UTC).

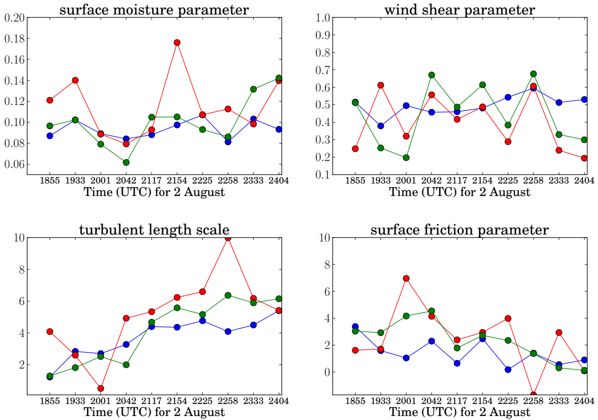
Parameter Estimation Experiments

- Matrix-free EnKF utilized for assimilation of data to obtain model parameters
- Three data assimilation experiments: assimilation of latent heat (DA1), horizontal winds (DA2), and both (DA3)
- Only observations were dBz is non-zero are assimilated (observation mask)
- Observations error is 16% for assimilation
- The 10 observational set are assimilated independently (no reinitialization)
- **Objective**: improvement in qualitative and quantitative structure of storm, and intensity

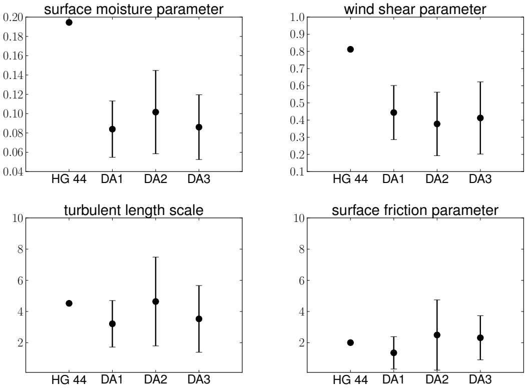
Before assimilating all 10 time periods of a given derived field, the impact of assimilating increasing amounts of observational data for a single time period was assessed.



EnKF parameter estimation as a function of number of latent heat observations assimilated at the first observational time period (1900 UTC). The additional latent heat observations were computed by adding vertical layers above and below the existing layers with the first layer being at 5 km in height.



Time distribution of the ensemble average parameter estimates with EnKF from DA1 (blue line, latent heat), DA2 (red line, horizontal winds), and DA3 (green line, both latent heat and horizontal winds).



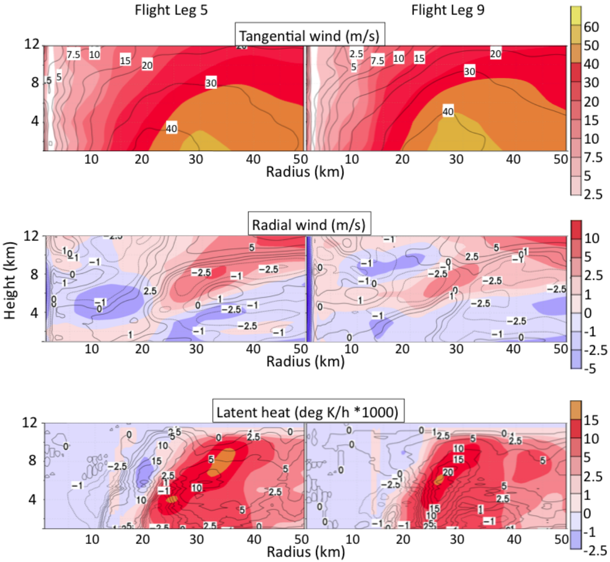
Analysis parameters averaged over time for ensemble member 44 (HG 44), DA1, DA2, and DA3. The vertical lines from the dots indicate the time variance of the parameter estimates for the various data sources.

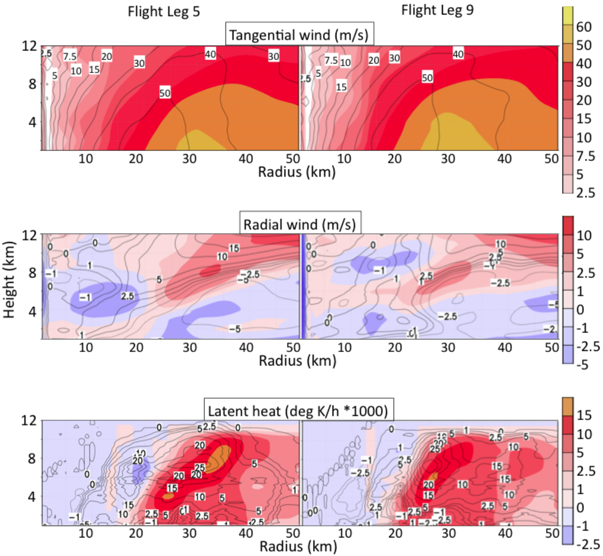
Simulation Experiments

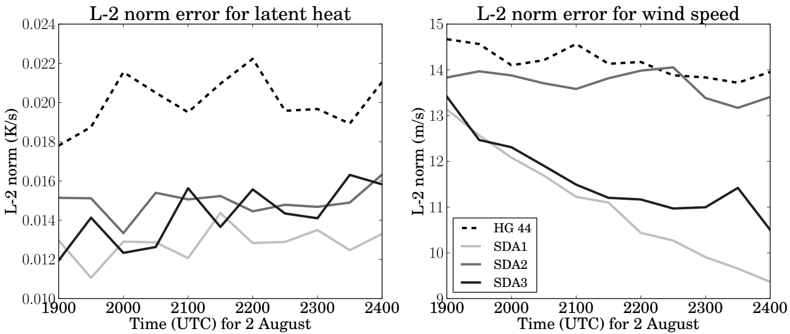
parameter/simulation or DA	DA1	DA2	DA3
surface moisture	0.08393103	0.10159057	0.0859986
wind shear	0.44383188	0.37766521	0.41229634
turbulent length scale	3.21374524	4.64259443	3.52878663
surface friction	1.35369145	2.49589051	2.31821474

Table: Time average parameter values for each of the data assimilation experiments DA1-DA3.

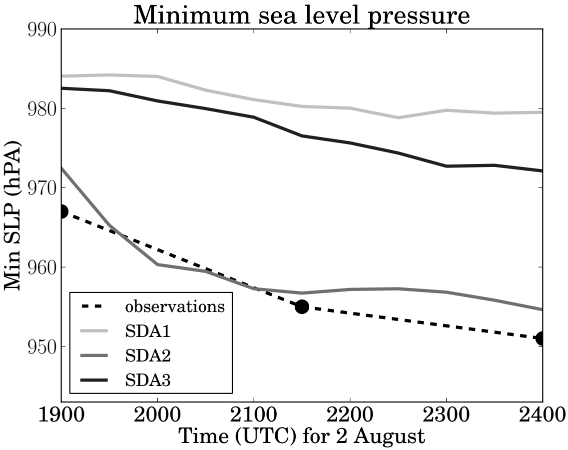
- Model simulations were performed with estimated time averaged parameters
- Three experiments were performed using parameters estimated using latent heat (SDA1), horizontal winds (SDA2), and both (SDA3)
- Estimated parameters are model and data dependent!


Flight Leg 5
Flight Leg 9
Tangential wind (m/s)
Radial wind (m/s)
Latent heat (deg K/h *1000)
Height (km)
Radius (km)


Flight Leg 5
Flight Leg 9
Tangential wind (m/s)
Radial wind (m/s)
Latent heat (deg K/h *1000)
Height (km)
Radius (km)
Radius (km)
Radius (km)



Error estimates as a function of time computed using 2-norm for ensemble member 44 (HG 44), SDA1, SDA2, and SDA3.



Minimum sea level pressure for SDA1, SDA2, and SDA3 along with the observed pressure from Hurricane Guillermo.

Conclusions

- Matrix-free EnKF more efficient for assimilation of large data sets than traditional matrix-oriented implementations
- Key model parameters are estimated through the assimilation of horizontal wind or latent heat fields available for Hurricane Guillermo
- Primary driver for a hurricane is latent heat release
- Depending on the derived data fields, the resulting parameters can produce a storm with the correct structure (latent heat) or the correct intensity (wind fields), observations near surface?
- Estimated parameters from different data are within one std of each other, differences in simulation suggest high sensitivity to their values
- Utilization of other data fields, such as radar reflectivity, require the model to faithfully capture physical processes that are not yet well understood

Future Work

- Localization in Matrix-free implementation (Khatri-Rao identity)
- In order for a given hurricane model to both reproduce the correct structure and intensity, numerical errors, especially near cloud edges, must be small
- Reduce the impact of cloud-edge errors either via the calibration of a tuning coefficient employed within an evaporative limiter (Reisner and Jeffery 2009,MWR)

Algorithm 1 Sherman-Morrison solver

procedure SHERMAN-MORRISON($\mathbf{A}_0$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{b}$, $\mathbf{x}$)
    Solve $\mathbf{A}_0\mathbf{x}_0 = \mathbf{b}$
    Solve $\mathbf{A}_0\mathbf{y}_{0,k} = \mathbf{u}_k$ for $k = 1, \ldots, N$
    for $i = 1, \ldots, N-1$ **do**
        $\mathbf{x}_i = \mathbf{x}_{i-1} - \frac{\mathbf{v}_i^T \mathbf{x}_{i-1}}{1+\mathbf{v}_i^T \mathbf{y}_{i-1,i}} \mathbf{y}_{i-1,i}$
        for $k = i+1, \ldots, N$ **do**
            $\mathbf{y}_{i,k} = \mathbf{y}_{i-1,k} - \frac{\mathbf{v}_i^T \mathbf{y}_{i-1,k}}{1+\mathbf{v}_i^T \mathbf{y}_{i-1,i}} \mathbf{y}_{i-1,i}$
        end for
    end for
    $\mathbf{x}_N = \mathbf{x}_{N-1} - \frac{\mathbf{v}_N^T \mathbf{x}_{N-1}}{1+\mathbf{v}_N^T \mathbf{y}_{N-1,N}} \mathbf{y}_{N-1,N}$
end procedure